# STAC67: Regression Analysis

Assignment 3
(Total: 100 points)

Please submit R Markdown file for Q. 4- Q. 5 along with your submission of the assignment.

Q.1 (24 points) Show the following statements.

(a) (4 pts) SSR (Sum of Squares of Regression) in matrix notation is:

$$\underset{\sim}{\widehat{\beta}}'\mathbf{X}'\underset{\sim}{Y} - \frac{1}{n}\underset{\sim}{Y}'\mathbf{J}\underset{\sim}{Y}$$

(b) (4 pts) Show that $\frac{1}{n}\mathbf{J}$, $\mathbf{H} - \frac{1}{n}\mathbf{J}$, and $\mathbf{I} - \mathbf{H}$ are idempotent and pairwise orthogonal (i.e. the product of each pair gives $\mathbf{0}$).

(c) (4 pts) Show that $\frac{SSR}{\sigma^2}$ is distributed as a non-central chisqure with $p' - 1$ degrees of freedom.

(d) (4 pts) Show that $\frac{SSE}{\sigma^2}$ is distributed as a $\chi^2_{n-p'}$ degrees of freedom

(e) (4 pts) Show that $\frac{SSR}{\sigma^2}$ and $\frac{SSE}{\sigma^2}$ are independent.

(f) (4 pts) We consider the general linear hypothesis test:

$$H_0 : \mathbf{K}'\underset{\sim}{\beta} = \underset{\sim}{m} \quad vs \quad H_a : \mathbf{K}'\underset{\sim}{\beta} \neq \underset{\sim}{m}$$

for a $p' \times k$ nonsingular matrix $\mathbf{K}$ and a $k \times 1$ vector $\underset{\sim}{m}$. Show that the F-test is a particular case of the general linear hypothesis test.

Q. 2 (10 points) A researcher fits a multiple linear regression model, relating yield (Y) of a chemical process to temperature ($X_1$), and the amounts of 2 additives ($X_2$ and $X_3$, respectively). She fits the following model:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

She wishes to test the following three hypotheses simultaneously:

- The mean response when $X_1 = 70,\ X_2 = 10, X_3 = 10$ is 80
- The average yield increases by 4 units when temperature increases by 1, controlling for $X_2$ and $X_3$
- The partial effect of increasing each additive is the same (controlling for all other factors)

(a) Specify following matrix and vectors that she is testing (this is her null hypothesis):

$$H_0 : \boldsymbol{K}'\underset{\sim}{\beta} - \underset{\sim}{m} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \Rightarrow$$

(b) She obtains the following results from fitting the regression based on $n = 24$ measurements while conducting the experiment:

$$(\boldsymbol{K}'\underset{\sim}{\widehat{\beta}} - \underset{\sim}{m})'(\boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K})^{-1}(\boldsymbol{K}'\underset{\sim}{\widehat{\beta}} - \underset{\sim}{m}) = 1800, \quad \underset{\sim}{Y}'(I - H)\underset{\sim}{Y} = 7800$$

Conduct her test at the $\alpha = 0.05$ significance level.

Q. 3 (20 points) Suppose that $X$ is a categorical variable with 3 levels (A, B, C) and we define the indicator variable $I_1$ and $I_2$ as:

$$I_1 = \begin{cases} 1, & X = A \\ 0, & \text{otherwise} \end{cases} \qquad I_2 = \begin{cases} 1, & X = B \\ 0, & \text{otherwise} \end{cases}$$

For a continuous response variable $Y$ consider fitting the linear model

$$Y = \beta_0 + \beta_1 I_1 + \beta_2 I_2 + \epsilon.$$

We take a total sample of $n$ individuals. Let $n_A$, $n_B$, $n_C$ be the number of individuals in each category of X and let $\bar{y}_A$, $\bar{y}_B$, $\bar{y}_C$ be the sample means of $Y$ for individuals in each category of X

(a) (5 pts) Find $\boldsymbol{X}'\boldsymbol{X}$ and $\boldsymbol{X}'\underset{\sim}{Y}$.

(b) (10 pts) Show that the least squares estimates for this model are

$$\hat{\beta}_0 = \bar{y}_C, \quad \hat{\beta}_1 = \bar{y}_A - \bar{y}_C, \quad \hat{\beta}_2 = \bar{y}_B - \bar{y}_C.$$

using both options (each option is 5 points each)
(option 1) $\hat{\beta} = (X^t X)^{-1} X\mathbf{y}$.
(option 2) For any parameter values $\beta_0, \beta_1, \beta_2$ we therefore need to minimize the sum of squared errors

$$S(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 I_{1i} - \beta_2 I_{2i})^2.$$

(c) (5 pts) Let $s_A^2$, $s_B^2$, $s_C^2$ be the usual sample standard deviations of $Y$ for individuals in each category of $X$. Show that the error sum of squares can be written as

$$SSE = (n_A - 1)s_A^2 + (n_B - 1)s_B^2 + (n_C - 1)s_C^2$$

Q. 4 (20 points) The public health department wished to study the relation between the average estimated probability of acquiring an infection in the hospital (**infections**, in percent; higher is worse) and the average length of stay of all patients in hospital (**StayLength** in days, $X_1$), the average age of patients (**Age**, in years, $X_2$), the average number of beds in hospital during study period (**Beds**, $X_3$). The data file, "Infectons.csv" can be found in Quercus. Please ignore the other three variables (MedSchool, Region,and Nurses) for this question.

(a) (4 pts) Obtain the scatter plot matrix and the correlation matrix. Interpret these and state your principal findings. Is there any concern about multicollinearity?

(b) (4 pts) Fit regression model for three predictor variables to the data and state the estimated regression function. How is $\hat{\beta}_2$ interpreted here?

(c) (4 pts) Test whether there is a regression relation; use $\alpha = 0.05$. State the alternatives, decision rule, and conclusion. What does your test imply about $\beta_1$, $\beta_2$, and $\beta_3$? What is the $P$-value of the test?

(d) (4 pts) Calculate the coefficient of determination, and also adjusted coefficient of determination. What does it indicate here?

(e) (4 pts) Obtain a 90 % prediction interval for a new hospital infection rate when StayLength = 10, Age = 45, and Beds = 150. Interpret your prediction interval.

Q. 5 (26 pts) We will use the same dataset, "Infections.csv" in Question 4 for this question. Following are the description of variables that will be used:

- Infections (Y): the average estimated probability of acquiring an infection in the hospital, in percent; higher is worse
- Beds: the average number of beds in hospital during study period
- Region: geographic region (NE = Northeast, NC = North Central, S = South, W = West)

(a) (8 pts) Write down the full model with the interaction terms. Fit the full model in R. Compute the estimated regression functions for geographic region and plot them.

(b)(4 pts) Test whether the slopes relating the average number of beds to infections are the same for each geographic region at the $\alpha = 0.05$, significance level.

(c) (2 pts) What model would you choose for this data? Justify your answer.

(d) (6 pts) For the model you chose in (c), check and comment on the standard assumptions for regression model.

(e) (6 pts) Look for the transformation of Y and/or X (=Beds). Fit the regression with the transformed variable(s) without interaction and comment whether this model fits better.